

Aspekte der Gestaltung und Aussagekraft epidemiologischer Studien am Beispiel der Deutschen Wirbelsäulenstudie

Tausch, Anja P.

Veröffentlichungsversion / Published Version
Sammelwerksbeitrag / collection article

Empfohlene Zitierung / Suggested Citation:

Tausch, A. P. (2014). Aspekte der Gestaltung und Aussagekraft epidemiologischer Studien am Beispiel der Deutschen Wirbelsäulenstudie. In V. Grosser, M. Schiltenswolf, & K.-D. Thomann (Hrsg.), *Berufskrankheit "Bandscheibenbedingte Erkrankungen der Lendenwirbelsäule" (BK 2108) : wissenschaftliche Grundlagen, sozialrechtliche Bewertung, Anleitung zur medizinischen Begutachtung, Perspektiven ; Handbuch für Gutachter, Berufsgenossenschaften, Rechtsvertreter, Anwälte und Gerichte* (S. 146-166). Frankfurt am Main: Referenz-Verl. <https://nbn-resolving.org/urn:nbn:de:0168-ssoar-46703-3>

Nutzungsbedingungen:

Dieser Text wird unter einer Deposit-Lizenz (Keine Weiterverbreitung - keine Bearbeitung) zur Verfügung gestellt. Gewährt wird ein nicht exklusives, nicht übertragbares, persönliches und beschränktes Recht auf Nutzung dieses Dokuments. Dieses Dokument ist ausschließlich für den persönlichen, nicht-kommerziellen Gebrauch bestimmt. Auf sämtlichen Kopien dieses Dokuments müssen alle Urheberrechtshinweise und sonstigen Hinweise auf gesetzlichen Schutz beibehalten werden. Sie dürfen dieses Dokument nicht in irgendeiner Weise abändern, noch dürfen Sie dieses Dokument für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, aufführen, vertreiben oder anderweitig nutzen.

Mit der Verwendung dieses Dokuments erkennen Sie die Nutzungsbedingungen an.

Terms of use:

This document is made available under Deposit Licence (No Redistribution - no modifications). We grant a non-exclusive, non-transferable, individual and limited right to using this document. This document is solely intended for your personal, non-commercial use. All of the copies of this documents must retain all copyright information and other information regarding legal protection. You are not allowed to alter this document in any way, to copy it for public or commercial purposes, to exhibit the document in public, to perform, distribute or otherwise use the document in public.

By using this particular document, you accept the above-stated conditions of use.

ANJA P. TAUSCH

Aspekte der Gestaltung und Aussagekraft
epidemiologischer Studien am Beispiel
der Deutschen Wirbelsäulenstudie

1 Einführung

Ärztliche Sachverständige, Mitarbeiter gesetzlicher Unfallversicherungen, Gewerkschafter, Rechtsbeistände oder Richter, die sich mit der Anerkennung von Erkrankungen als Berufskrankheiten beschäftigen, stehen immer wieder vor der Herausforderung, sich mit aktuellen wissenschaftlichen Erkenntnissen auseinandersetzen zu müssen. Hauptinformationsquelle sind hierbei epidemiologische Studien, die sich mit der Häufigkeit des Auftretens von Krankheiten und deren Risikofaktoren in verschiedenen Gruppen der Bevölkerung beschäftigen. Immer wieder ergibt sich dabei das Problem, mit unterschiedlichen, sich teilweise widersprechenden Ergebnissen konfrontiert zu sein, deren Aussagekraft und Anwendungsbereich beurteilt werden muss. Nicht zuletzt hängen diese von der Art des gewählten Studiendesigns ab, wobei nicht unbedingt prinzipiell von „besseren“ bzw. „schlechteren“, sondern höchstens von „dem Gegenstand oder der Fragestellung angemesseneren“ bzw. „weniger angemessenen“ Designs gesprochen werden kann. Darüber hinaus ist es nicht ganz unbedeutend für die Beurteilung der Verallgemeinerbarkeit der Ergebnisse einer Studie, wie mit den jeweiligen – häufig studiendesignspezifischen – unbeabsichtigten Störungen umgegangen

wurde (z.B. ob diese kontrolliert und/oder bei der Interpretation berücksichtigt wurden). Der folgende Beitrag soll eine Hilfestellung dazu leisten, die Vielzahl von Designs epidemiologischer Studien, deren Aussagekraft und Angemessenheit für einen Untersuchungsgegenstand und die Vergleichbarkeit von Ergebnissen besser einordnen und beurteilen zu können.

Dazu werden zunächst die wichtigsten Studiendesigns mit ihren Vor- und Nachteilen sowie mögliche unbeabsichtigte Einflüsse (Störgrößen) dargestellt. Dieser Überblick soll dabei helfen, *einzelne* Studien und ihre Ergebnisse besser beurteilen und vergleichen zu können. Die von Bradford Hill (1965) vorgestellten Kriterien, auf welche im Anschluss eingegangen wird, geben Anhaltspunkte dafür, wie der *gesamte* Forschungsstand zum Zusammenhang eines Risikofaktors mit einer Erkrankung zu beurteilen ist. Abschließend sollen Anregungen gegeben werden, wie die zuvor dargestellten Informationen dazu herangezogen werden können, das Design und die Ergebnisse der Deutschen Wirbelsäulenstudie (u. a. Ellegast et al. 2007; Linhardt et al. 2007) einzuordnen und damit die Stärken, aber auch eventuelle kritische Punkte dieser Studie besser einschätzen zu können.

2 Grundbegriffe epidemiologischer Forschung

Wie bereits angesprochen, beschäftigt sich epidemiologische Forschung mit der Häufigkeit, Verteilung und den Ursachen von Krankheiten. Generell dient dabei als Datengrundlage das „Zählen klinischer Ereignisse in Gruppen ähnlicher Fälle“ (Fletcher & Fletcher 2007: 22). Dabei werden als *klinische Ereignisse* meist Krankheiten (z.B. Herzinfarkt), krankheitsbezogene Ereignisse (Reinfarkte) oder Todesfälle („Mortalität“), als *Gruppen ähnlicher Fälle* meist Personen mit einem bestimmten Risikofaktor („Exponierte“) bzw. ohne diesen Risikofaktor („Nicht-Exponierte“) untersucht. Welche Methoden hierbei im Einzelnen unterschieden werden können, darauf wird im nächsten Abschnitt eingegangen. Zunächst sollen aber drei wichtige Begriffspaare kurz definiert und einander gegenübergestellt werden.

Für die Beschreibung der Häufigkeit des Auftretens von Erkrankungen in der Bevölkerung werden zunächst die Begriffe *Prävalenz* und *Inzidenz* verwendet. Während Prävalenz den Anteil von Personen in einer bestimmten Bevölkerungsgruppe beschreibt, der zu einem bestimmten Zeitpunkt (Punktprevalenz) oder in einem bestimmten Zeitraum (Periodenprävalenz) an einer spezifischen Krankheit *leidet*, werden für die Inzidenz nur solche Personen gezählt, die in einem definierten Zeitraum *neu* erkranken. Je länger eine Erkrankung andauert, um so stärker unterscheiden sich Prävalenz und Inzidenz, wobei der Wert der Prävalenz naturgemäß immer höher ist als der der Inzidenz. (Zum Zusammenhang zwischen Prävalenz, Inzidenz und Erkrankungsdauer vgl. auch Fletcher & Fletcher 2007: 94-95.)

Für die Beschreibung von Studienarten und dabei im Speziellen der Art der Erhebung des Zusammenhangs zwischen Risikofaktor und Krankheit werden weiterhin die Begriffe *retrospektiv* und *prospektiv* verwendet (vgl. Abbildung 1). Von einer retrospektiven Studie wird gesprochen, wenn die Krankheit zum Untersuchungszeitpunkt bereits di-

agnostiziert wurde. In einem solchen Fall wird der Risikofaktor – und somit auch sein Zusammenhang mit der Krankheit – „rückblickend“ (retrospektiv) erhoben. Prospektive Studien sind dagegen auf die Zukunft ausgerichtet und erfordern mindestens zwei Untersuchungszeitpunkte. Diese werden so gewählt, dass der erste der Erhebung des Risikofaktors dient, wobei die entsprechende Krankheit noch nicht aufgetreten sein darf. Deren Vorhandensein wird dann erst zum zweiten Zeitpunkt registriert und mit dem Risikofaktor in Verbindung gebracht. Die in Abbildung 1 wiedergegebene Grafik von Röhrig et al. (2009) stellt die Erhebungsrichtungen prospektiver (a) und retrospektiver (b) Studien einander gegenüber.

Zwei weitere Begriffe begegnen dem Leser in der epidemiologischen Forschung häufig: Dies sind

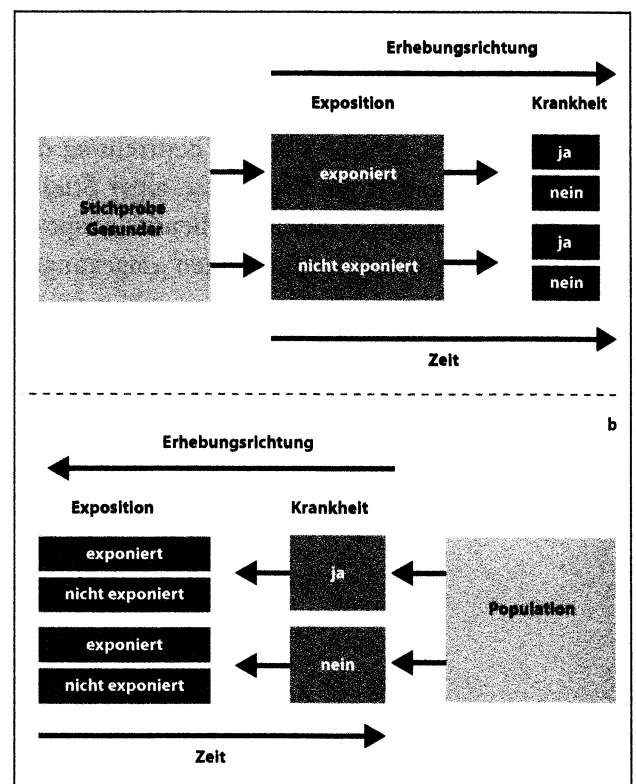


Abbildung 1:
Vergleichende Darstellung der Erhebungsrichtung prospektiver (oben) und retrospektiver Studien (unten; aus Röhrig et al. 2009: 264)

		Risikofaktor		
		Ja „Exponierte“ <i>(z.B. Raucher)</i>	Nein „Nicht-Exponierte“ <i>(Nicht-Raucher)</i>	
Erkrankung	Ja: „Fälle“ <i>(z.B. Lungenkrebs)</i>	25 (A1)	16 (A2)	41 (A1+A2)
	Nein: „Kontrollen“ <i>(kein Lungenkrebs)</i>	369 (B1)	1591 (B2)	1960 (B1+B2)
		394 (A1+B1)	1607 (A2+B2)	2001 (A1+A2+B1+B2)

Tabelle 1:
Beispielhafte Häufigkeitstabelle zur Veranschaulichung der Berechnung von Relativem Risiko und Odds Ratio (Erläuterung vgl. Text)

das *Relative Risiko (RR)* und das *Odds Ratio (OR)*. Beiden liegt eine Vierfeldertafel der Häufigkeiten von Personen mit den Merkmalskombinationen Risikofaktor Ja/Nein und Erkrankung Ja/Nein zugrunde (vgl. Tabelle 1). Es werden jedoch jeweils unterschiedliche Personengruppen zueinander ins Verhältnis gesetzt.

Das Relative Risiko (RR) beantwortet die Frage, wie viel mal höher das Erkrankungsrisiko der Gruppe mit Risikofaktor („Exponierte“; Erkrankungsrisiko = Raucher mit Lungenkrebs im Verhältnis zu allen Rauchern) im Vergleich zur Gruppe ohne Risikofaktor („Nicht-Exponierte“; Erkrankungsrisiko = Nicht-Raucher mit Lungenkrebs im Verhältnis zu allen Nicht-Rauchern) ist. Mit Bezug auf die (fiktiven) Zahlen in Abbildung 1 bedeutete dies: $RR = (A1 / (A1 + B1)) / (A2 / (A2 + B2))$ oder $(25 / 394) / (16 / 1607) = 6,37$, d.h. für einen Raucher wäre das Risiko, Lungenkrebs zu bekommen 6,37-mal so hoch wie für einen Nicht-Raucher. Es handelt sich dabei also um ein Maß, das dazu dient, aus dem Vorhandensein eines Risikofaktors ein zukünftiges Erkrankungsrisiko vorherzusagen, also um ein *prospektives Maß*.

Während für das Relative Risiko die beiden Gruppen mit und ohne Risikofaktor (Gruppe 1 und 2) verglichen werden, beruht das Odds Ratio auf dem Vergleich der Erkrankten mit den Nicht-Erkrankten (Gruppe A und B). Hierfür wird das Verhältnis der Erkrankten mit und ohne Risikofaktor in Beziehung zu dem entsprechenden Verhältnis der Nicht-Er-

krankten gesetzt. Mit Bezug auf die Zahlen in Abb. 1 bedeutete dies: $OR = (A1 / A2) / (B1 / B2)$ oder $(25 / 16) / (369 / 1591) = 6,74$, d.h. in der Gruppe der Lungenkrebskranken sind 6,74-mal so viele Personen Raucher wie in der Gruppe der Nicht-Erkrankten. Da das Odds Ratio sozusagen von der bereits eingetretenen Erkrankung auf den Risikofaktor „zurückblickt“, ist es ein *retrospektives Maß*.

Ungeachtet der unterschiedlichen Berechnung wird auch das OR häufig als ein Maß des Erkrankungsrisikos interpretiert. Tatsächlich ist es aber so, dass nur für Erkrankungen mit einer geringen Auftretenswahrscheinlichkeit in der Bevölkerung RR und OR – wie in dem Beispiel – etwa gleich hoch sind. (Fletcher & Fletcher 2007: 139 geben hierzu als Grenze eine Inzidenz bei Nicht-Exponierten von < 1% an.) Je höher die Erkrankungsrate, um so größer ist das OR dagegen im Vergleich zum RR. Insgesamt ist das Relative Risiko etwas anschaulicher und leichter nachzuvollziehen. Es kann jedoch nur in solchen Studien berechnet werden, in denen die Randwahrscheinlichkeiten der Häufigkeitstabelle zufällig sind, d.h. in denen die Anzahl der untersuchten Erkrankten und Nicht-Erkrankten nicht durch das Studiendesign fest vorgegeben ist. Bei dem für die Deutsche Wirbelsäulenstudie verwendeten Design der Fall-Kontroll-Studie, bei welchem die gleiche Anzahl von Erkrankten und Nicht-Erkrankten untersucht wird (vgl. folgenden Abschnitt), ist die Berechnung des Relativen Risikos also nicht möglich.

3 Studienarten

Abbildung 2 zeigt einen Überblick über die wichtigsten Untersuchungsdesigns epidemiologischer Studien. Grundsätzlich können dabei beobachtende (nicht interventionelle) und experimentelle (interventionelle) Studien unterschieden werden. Beobachtende Studien gewinnen ihre Daten allein durch die Registrierung der Häufigkeit des Auftretens bestimmter Faktoren und Erkrankungen. Sie beobachten damit nur eine Beziehung zwischen Risikofaktoren und Erkrankungen, greifen in diese Beziehung jedoch nicht ein. Im Gegensatz dazu erfordern experimentelle Studien immer eine Intervention, d. h. eine absichtliche Beeinflussung der vermuteten Ursachefaktoren sowie die Beobachtung der auftretenden Veränderungen. Bei der Mehrzahl der epidemiologischen Studien handelt es sich um Beobachtungsstudien, Interventionsstudien machen dagegen, v. a. wegen häufiger ethischer Schwierigkeiten, nur einen geringen Anteil aus.

In Abbildung 2 sind die Studiendesigns so angeordnet, dass ihre Aussagekraft zu der Frage „Ist ein bestimmter Faktor die Ursache einer Erkrankung?“ von links nach rechts zunimmt. Gleichzeitig wächst jedoch auch der Aufwand, der für eine bestimmte Studienart erforderlich ist. Somit hat unter Berücksichtigung ökonomischer Gesichtspunkte und abhängig vom jeweiligen Forschungsstand und der Forschungsfrage jedes Design seine Berechtigung. Liegen beispielsweise noch keinerlei Vermutungen über mögliche Ursachen einer Erkrankung vor, können bereits die relativ leicht zugänglichen Daten einer ökologischen Studie erste wichtige Hinweise für prinzipiell infrage kommende Risikofaktoren liefern. Im Folgenden werden die fünf in Abbildung 2 dargestellten Studiendesigns etwas ausführlicher beschrieben und ihre Vor- und Nachteile diskutiert.

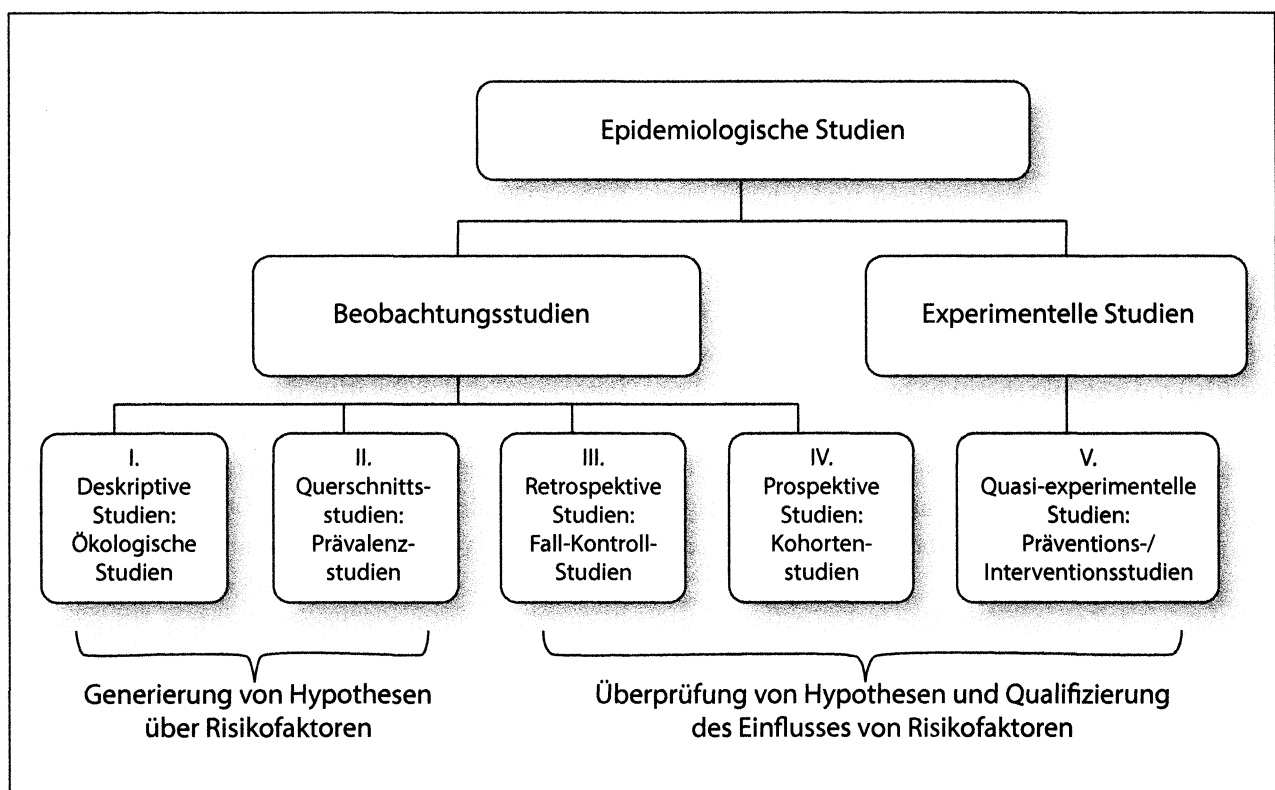


Abbildung 2:
Überblick über die wichtigsten Studiendesigns epidemiologischer Studien

3.1 Ökologische Studien

Bei der Durchführung Ökologischer Studien werden selbst keine Daten erhoben, sondern es wird auf bereits vorliegende statistische Werte der Verteilungen von Krankheiten und Risikofaktoren in bestimmten Bevölkerungsgruppen zurückgegriffen. Ökologische Studien dienen der Darstellung von statistischen Zusammenhängen innerhalb von Bevölkerungsgruppen und sind daher deskriptiv (beschreibend). Gleichzeitig sind ihre Analyseeinheit nicht – wie in den anderen Studientypen – Individuen, sondern sogenannte Aggregate, d.h. Gruppen ähnlicher Individuen, wie z. B. Einwohner verschiedener Länder. Bei der Analyse werden die Häufigkeiten von Erkrankungen (oder auch Sterberaten) der verschiedenen Bevölkerungsgruppen in Beziehung gesetzt zu deren Häufigkeiten eines Risikofaktors. Als statistische Maße werden dabei Korrelationskoeffizienten¹ berechnet. So zeigt die in Abbildung 3 wiedergegebene Grafik aus einer Studie von Wolf-Maier et al. (2003: 2367) den Zusammenhang zwischen der Sterblichkeitsrate für Herzinfarkt und der Prävalenz von Bluthochdruck in verschiedenen Ländern. Der in der Studie ermittelte Korrelationskoeffizient von $r = .78$ weist auf einen relativ hohen positiven Zusammenhang zwischen den beiden Variablen hin, was sich in Abbildung 3 auch darin zeigt, dass sich die Markierungen der einzelnen Länder relativ nah an einer gedachten Diagonale anordnen.

Ökologische Studien verursachen relativ geringe Kosten. Ihre Ergebnisse liefern jedoch nur Hinweise auf *Zusammenhänge* zwischen Risikofaktoren

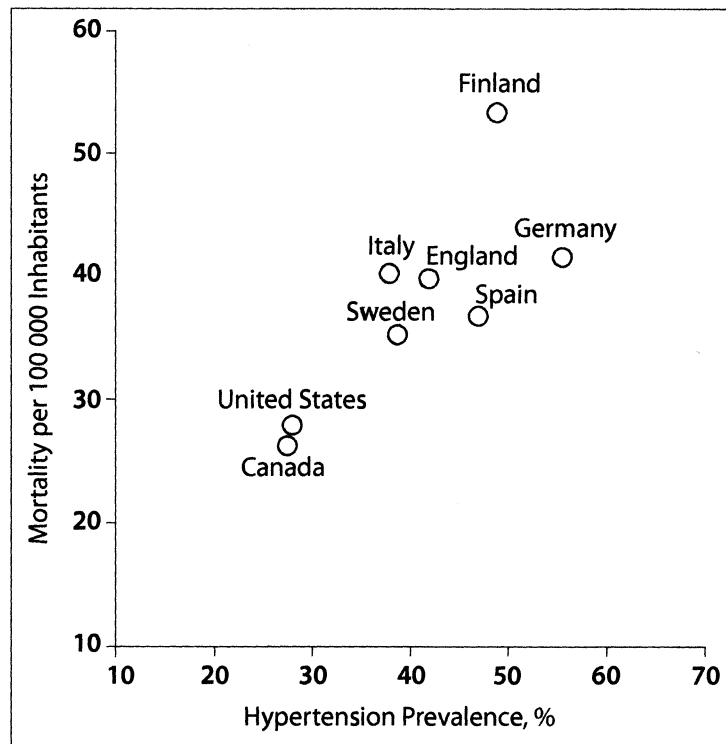


Abbildung 3:
Zusammenhang zwischen der Prävalenz von Bluthochdruck und der Sterblichkeit in verschiedenen Europäischen und Nord-Amerikanischen Staaten (aus Wolf-Maier et al. 2003)

und Krankheiten. Sie erlauben jedoch keinerlei Aussagen darüber, ob die Risikofaktoren tatsächlich als *Ursachen* für die Krankheit angesehen werden können. Darüber hinaus können Störfaktoren (vgl. Abschnitt 4) in diesem Studiendesign nicht oder nur schwer kontrolliert werden.

3.2 Querschnittsstudien: Prävalenzstudien

Querschnitts- oder Prävalenzstudien untersuchen ebenfalls die Zusammenhänge der Häufigkeit einer Erkrankung in der Bevölkerung (Inzidenz/Prävalenz) mit ihren vermuteten Risikofaktoren. Im Gegensatz zu Ökologischen Studien werden diese Zusammenhänge jedoch nicht auf der Ebene von Personengruppen, sondern auf der Ebene von Individuen analysiert. Dies erfordert somit auch eine der Fragestellung entsprechende Datenerhebung. Auch bei dieser Studienart besteht der größte Nachteil

¹ Maß für den Zusammenhang zwischen zwei Variablen mit einem Wertebereich zwischen -1 und 1. Je höher der Betrag des Wertes, desto stärker ist der Zusammenhang, wobei 1 einen perfekten proportionalen (je höher die eine Variable, um so höher die andere Variable), -1 einen perfekten antiproportionalen (je höher die eine Variable, um so niedriger die andere Variable) und 0 einen nicht vorhandenen Zusammenhang anzeigt.

darin, dass Risikofaktor(en) und Krankheit zum selben Zeitpunkt erhoben werden und somit keine Aussagen über Ursache und Wirkung möglich sind. Die Kontrolle von Störfaktoren ist hier jedoch eher möglich, da infrage kommende Variablen bei der Datenerhebung berücksichtigt werden können.

3.3 Retrospektive Studien: Fall-Kontroll-Studien

Obwohl auch Fall-Kontroll-Studien nur auf einem Erhebungszeitpunkt beruhen, werden hierbei durch die Art der Datenerhebung Aussagen über die zeitliche Abfolge von Risikofaktor und Krankheiten möglich. Dazu wird eine Gruppe Erkrankter über bestimmte Risikofaktoren befragt, die bereits vor der Erkrankung aufgetreten sind. Diese Personen werden mit einer etwa gleich großen und hinsichtlich demographischer Variablen (z. B. Alter, Geschlecht, Wohnort) ähnlichen Gruppe Nicht-Erkrankter verglichen. Da bei diesem Design die Anzahl der untersuchten Erkrankten und Nicht-Erkrankten gleich groß ist, muss bei der Abschätzung des Erkrankungsrisikos auf das statistische Maß des Odds Ratio zurückgegriffen werden. Genau genommen wird hier also die Frage untersucht, wie wahrscheinlich das Vorhandensein bestimmter Risikofaktoren in der Gruppe der Erkrankten im Vergleich zur Gruppe der Nicht-Erkrankten ist (vgl. Tabelle 2).

Da die Risikofaktoren zum Zeitpunkt der bereits diagnostizierten Erkrankung rückblickend erhoben werden, handelt es sich bei Fall-Kontroll-Studien um retrospektive Studien. Ein wesentlicher

Nachteil dieser Studien ist die Möglichkeit, dass Risikofaktoren rückblickend verändert wahrgenommen und damit die Ergebnisse verzerrt werden (Erinnerungsbias). Auch bei der Auswahl der Studienteilnehmer können Verzerrungen auftreten, wenn ganz bestimmte Personen in der Gruppe der Erkrankten oder der Nicht-Erkrankten die Teilnahme verweigern (selektive Drop Outs). Auf diese beiden Punkte wird im folgenden Abschnitt über die Störfaktoren noch einmal ausführlicher eingegangen. Als großer Vorteil der Fall-Kontroll-Studien kann die Möglichkeit angesehen werden, Risikofaktoren theoriegeleitet auszuwählen und dabei auch den kombinierten Einfluss verschiedener Risikofaktoren (multiple Exposition) untersuchen zu können. Um die genannten Nachteile zu minimieren und möglichst aussagekräftige Ergebnisse zu erhalten, sollten verschiedene Regeln beachtet werden:

Erstens sollten niemals nur Patienten einer Praxis oder eines Krankenhauses befragt werden. Je mehr Untersuchungsorte gewählt werden („multizentrische Studien“), um so eher lassen sich selektive Drop Outs minimieren. Zweitens sollten für die Patientenstichprobe nur Neu-Erkrankte ausgewählt werden. Dies erhöht die Wahrscheinlichkeit, dass die genannten Risikofaktoren bereits vor der Erkrankung bestanden. Drittens sollte die nicht erkrankte Kontrollgruppe der Patientenstichprobe möglichst ähnlich sein, d. h. aus derselben Region stammen, einen vergleichbaren Sozialstatus besitzen und die gleiche Alters- und Geschlechtsverteilung aufweisen. Ideal ist ein sogenanntes Matching, bei dem zu jedem untersuchten Patienten eine nicht erkrankte Person gesucht wird, die ihm hinsichtlich

	Lungenkrebs („Fälle“)	Kein Lungenkrebs („Kontrolle“)	
Rauchen „ja“	4.932	3.769	8.728
Rauchen „nie“	375	1.659	2.034
	5.307	5.455	10.762
Verhältnis mit RF/ohne RF	13,15:1	2,27:1	OR \approx 5,8

Tabelle 2:
Häufigkeitstabelle der Verteilung des Risikofaktors (RF) Rauchen und der Erkrankung Lungenkrebs (Dt. Ärzteblatt, 1997)

der genannten Merkmale möglichst ähnlich ist. Besondere Sorgfalt muss darüber hinaus der Erfassung der Risikofaktoren gewidmet werden. Werden diese in Patienten- und Kontrollgruppe von unterschiedlichen Personen erhoben, können Verzerrungen durch unterschiedliche Arten der Befragung auftreten. Auf jeden Fall sollte den Untersuchern nicht bekannt sein, ob die jeweilige befragte Person der Patienten- oder der Kontrollgruppe angehört. Ansonsten bestünde die Gefahr, dass Erkrankte im Sinne der Fragestellung anders befragt würden als Nicht-Erkrankte (z.B. durch nachdrücklicheres Nachfragen) und somit der Einfluss der Risikofaktoren überschätzt würde.

3.4 Prospektive Studien: Kohortenstudien

Bei dieser Studienart handelt es sich nun um das erste Design, das (mindestens) zwei Untersuchungszeitpunkte umfasst. Abbildung 4 stellt den Ablauf von Kohortenstudien dar: Die erste Erhebung (Messzeitpunkt 1) findet zu einem Zeitpunkt statt,

zu dem die Bevölkerung (Population) noch keine Erkrankungen zeigt. Da nicht die gesamte Population untersucht werden kann, wird eine Zufallsstichprobe ausgewählt und in dieser das Vorkommen bestimmter Risikofaktoren (z. B. Hypertonie) erhoben. Erst zu einem späteren Zeitpunkt wird erfasst, wie viele bzw. welche Personen aus der Stichprobe eine bestimmte Krankheit entwickelt haben. Damit werden einige kritische Punkte der zuvor beschriebenen Fall-Kontroll-Studien vermieden: Da zum Zeitpunkt der Erhebung der Risikofaktoren alle Befragten noch nicht erkrankt sind, wird eine systematische Ungleichbehandlung bei der Befragung der (zukünftigen) Patientenstichprobe und der Kontrollstichprobe ausgeschlossen. Darüber hinaus kann eindeutig sichergestellt werden, dass die Risikofaktoren schon vor Beginn der Erkrankung bestanden, womit erstmals eindeutige kausale Rückschlüsse möglich sind.

Die wesentlichen Nachteile von Kohortenstudien bestehen in ihrem hohen zeitlichen und personalen Aufwand. Neben der sich über mehrere Jahre erstreckenden Untersuchungsdauer muss auch die Anfangsstichprobe – abhängig von der

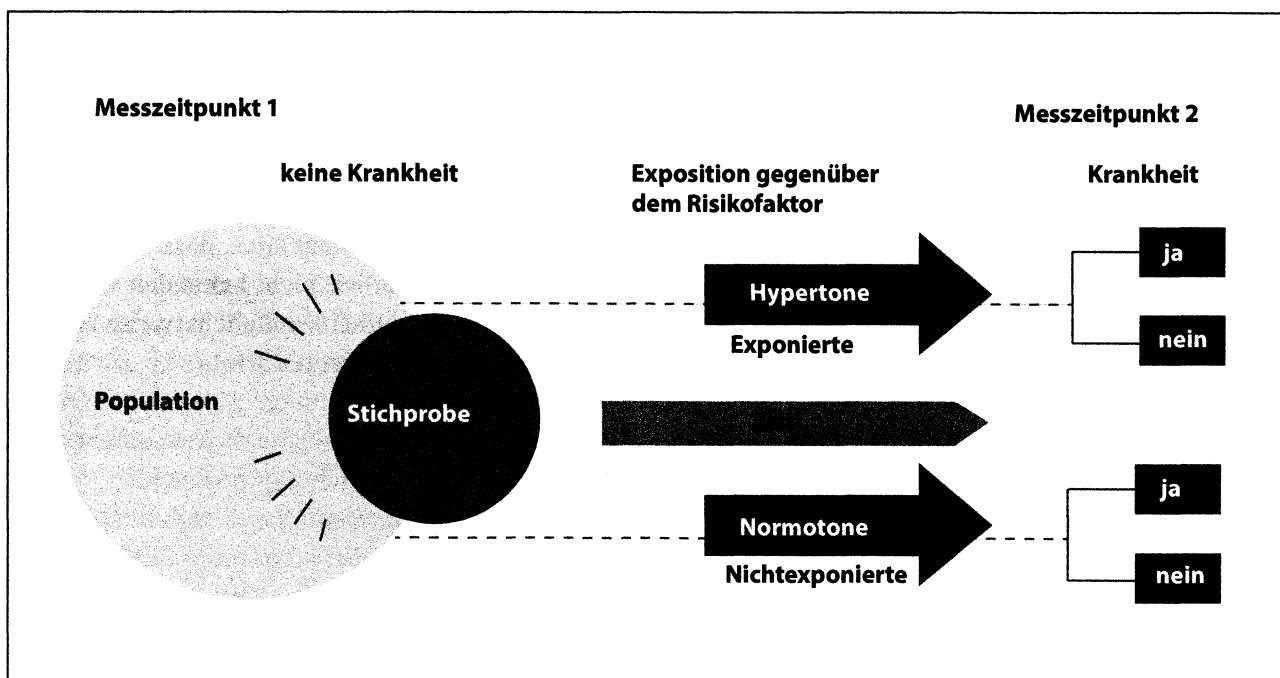


Abbildung 4:
Ablauf von Kohortenstudien am Beispiel des Risikofaktors Hypertonie (nach Fletcher & Fletcher 2007)

Prävalenzrate der Erkrankung – ausreichend groß gewählt werden. Nur dann ist die spätere Gruppe der Erkrankten groß genug, um statistisch haltbare Aussagen über den Zusammenhang zwischen Risikofaktor und Erkrankung machen zu können. Das hierbei verwendete Maß ist das Relative Risiko, mit dem das Erkrankungsrisiko von exponierten und nicht exponierten Personen verglichen wird.

3.5 Quasi-experimentelle Studien: Präventions-/Interventionsstudien

Die stärkste Aussagekraft im Hinblick auf die Kausalität von Risikofaktoren für bestimmte Erkrankungen haben Präventions- oder Interventionsstudien. Wie bereits beschrieben, greifen diese experimentell in den Zusammenhang zwischen Risikofaktor und Erkrankung ein, indem durch spezielle Maßnahmen Risikofaktoren reduziert oder beseitigt werden und die Auswirkungen dieser Maßnahmen auf das Auftreten (Präventionsstudie) bzw. den Verlauf (Interventionsstudie) der betrachteten Krankheit registriert werden. Während bei Kohortenstudien zwar sichergestellt ist, dass die betrachteten Risikofaktoren bereits vor der Erkrankung bestanden haben, kann die Entstehung der späteren Krankheit dennoch nicht immer eindeutig ursächlich auf die Risikofaktoren zurückgeführt werden. Theoretisch ist immer noch denkbar, dass in der zwischenzeitlich vergangenen Zeitspanne andere Faktoren wirksam waren, welche die Krankheit verursacht haben. Demgegenüber lassen Präventions- und Interventionsstudien eindeutige

Rückschlüsse darauf zu, dass eine Verbesserung des Krankheitsverlaufs oder eine Reduktion der Erkrankungsrate auch tatsächlich auf die Reduktion der Risikofaktoren zurückzuführen ist.

Auch bei dieser Studienart sind zwei Messzeitpunkte erforderlich (vor und nach der Maßnahme). Da diese aber meist nur einige Wochen oder Monate auseinanderliegen, ist der zeitliche Aufwand nicht ganz so groß wie bei den zuvor beschriebenen Kohortenstudien. Auch können sie aussagekräftige Ergebnisse bereits mit einer relativ geringen Zahl an untersuchten Personen erreichen. Wesentlich aufwändiger ist dagegen die zwischen den beiden Messzeitpunkten durchzuführende Maßnahme. Während die Zeit zwischen zwei Erhebungszeitpunkten einer Kohortenstudie nur mit Warten zu füllen ist, erfordern Präventions- und Interventionsstudien auch innerhalb des Messintervalls hohe Ressourcen. Der größte Nachteil dieser Studien liegt jedoch in den mit dem Design verbundenen ethischen Problemen. Um die Wirksamkeit der Intervention eindeutig feststellen zu können, reicht ein einfacher Vergleich zwischen dem Erkrankungsstatus der behandelten Personengruppe zum ersten und zweiten Messzeitpunkt nicht aus. Der Einfluss zufälliger, interventionsunabhängiger Veränderungen kann nur dann weitgehend kontrolliert werden, wenn die Effekte der Interventionsgruppe mit etwaigen Veränderungen bei einer nicht behandelten Kontrollgruppe verglichen werden. Somit beinhaltet dieses Studiendesign immer auch, dass eine Gruppe von Personen absichtlich nicht behandelt wird und diesen Patienten damit eine möglicherweise wirksame Intervention vorenthalten wird.

4 Störfaktoren

Die vorherigen Abschnitte machten bereits deutlich, dass jedes epidemiologische Studiendesign sowohl Vor- als auch Nachteile hat. Unabhängig von der gewählten Studienart gilt es also, die jeweiligen Vorteile möglichst gut zu nutzen und die Nachteile so gut wie möglich zu reduzieren. Für letzteres ist es notwendig, sich möglicher Störfaktoren bewusst zu sein, die je nach Design unterschiedlich stark wirken und in unterschiedlichem Maße kontrolliert werden können. Da es bei den meisten Designs nicht möglich ist, Störfaktoren vollständig auszuschalten, sollten diese auch bei der Interpretation und Beurteilung von Studienergebnissen immer im Hinterkopf behalten und als alternative Erklärungsmöglichkeiten in Betracht gezogen werden. Daher soll im folgenden Abschnitt auf verschiedene Gruppen von Störfaktoren, ihre Ursachen und mögliche Folgen etwas ausführlicher eingegangen werden.

Prinzipiell gilt, dass jede Studie vielfältigen Einflüssen ausgesetzt ist, die nicht vollständig durch das Untersuchungsdesign erfasst und kontrolliert werden können. Wirken diese auf Patienten- und Kontrollgruppe gleichermaßen, so ist ihr Einfluss bei der Interpretation zu vernachlässigen. Problematisch sind dagegen *systematische* Verzerrungen (engl. „Bias“), die auf eine der Gruppen stärker wirken als auf die andere und dadurch Effekte der Einflussvariablen (Risikofaktoren) vortäuschen oder aber auch geringer erscheinen lassen können (vgl. Fletcher & Fletcher 2007: 26 – 30). Im Wesentlichen können drei Arten von Verzerrungen auftreten: Auswahlverzerrungen (*Selection Bias*), Informationsverzerrungen (*Information Bias*) sowie weitere Störgrößen (*Confounding*). Ihre Ursachen sowie Maßnahmen, mit denen sie auf ein möglichst geringes Maß reduziert werden können, werden im Folgenden dargestellt.

4.1 Selection Bias (Auswahlverzerrung)

Eine Auswahlverzerrung liegt vor, wenn bestimmte Personen in der Patienten- gegenüber der Kontrollstichprobe über- oder unterrepräsentiert sind. Diese kommt dadurch zustande, dass bestimmte Personen aus der Patientenpopulation tendenziell nicht teilnehmen (*Non Response*) oder vermehrt teilnehmen (*Freiwilligen Bias*). Anhand des Beispiels einer Untersuchung über die Wirksamkeit einer technischen Neuerung bei der Behandlung von Krebserkrankungen können diese beiden Verzerrungsarten verdeutlicht werden: Einerseits sind Männer generell mehr an technischen Neuerungen interessiert und erklären sich daher möglicherweise eher freiwillig zur Teilnahme bereit (*Freiwilligen Bias*). Andererseits sind Frauen häufiger neben ihrer Erkrankung mit familiären Aufgaben belastet und haben daher weniger freie Kapazitäten für die Teilnahme an einer Studie (mehr *Non Response*). In diesem Fall würden beide Arten der Auswahlverzerrung dazu führen, dass in der endgültigen Stichprobe Frauen unter- und Männer überrepräsentiert wären. Ergebnisse dieser Studie könnten daher nur mit Vorbehalt auf weibliche Patienten übertragen werden.

Bei Studien mit mehreren Messzeitpunkten kann es darüber hinaus passieren, dass eine ganz bestimmte Teilgruppe von Patienten spätere Befragungstermine nicht mehr wahrnimmt (*selektive Drop Outs*). Gründe hierfür könnten lange Anfahrtswege sein (bei späteren Zeitpunkten sind dann möglicherweise in der Stadt lebende Patienten gegenüber auf dem Land lebenden Patienten überrepräsentiert) oder auch das vorzeitige Versterben (bei späteren Zeitpunkten sind dann möglicherweise leichtere Fälle gegenüber schwereren überrepräsentiert).

Wissenschaftlich aussagekräftige Studien sollten daher neben den im Zentrum stehenden Risikofaktoren immer auch weitere Merkmale erfassen, die mit der untersuchten Krankheit in Verbindung stehen könnten. Sie sollten darauf abzielen, Unter-

schiede zwischen den untersuchten Gruppen möglichst zu reduzieren oder – wenn dies nicht möglich ist – zumindest Informationen über die Verteilung relevanter Merkmale in den Gruppen zu liefern.

4.2 Informationsverzerrung (Information Bias)

Eine Informationsverzerrung tritt auf, wenn die relevanten Informationen über Risikofaktoren, Erkrankungen und weitere relevante Merkmale in Patienten- und Kontrollgruppe auf unterschiedliche Art und Weise erhoben werden. Dies kann der Fall sein, wenn Patienten- und Kontrollgruppe von unterschiedlichen Interviewern befragt werden (*Interviewer Bias*). Hierbei können sich personenspezifische Eigenheiten in der Art der Befragung auswirken, aber auch Unterschiede, die darauf zurückzuführen sind, dass die Interviewer Erkrankte anders befragen als Nicht-Erkrankte. So könnten sie beispielsweise bei ersteren nachdrücklicher in Richtung Risikofaktoren nachfragen oder ihre Befragung anschaulicher durch Beispiele untermalen und so eine stärkere Zustimmung zum Vorhandensein der untersuchten Risikofaktoren erreichen. Eine gute Studie wird daher „blind“ durchgeführt, d.h. dieselben Interviewer befragen sowohl Erkrankte als auch Nicht-Erkrankte, wobei sie keine Informationen darüber haben, welcher der beiden Untersuchungsgruppen ihr jeweiliges Gegenüber angehört.

Zum anderen können auch auf Seiten der Befragten Informationsverzerrungen auftreten. Es besteht die Möglichkeit, dass Erkrankte frühere Erfahrungen mit Risikofaktoren rückblickend anders erinnern als Nicht-Erkrankte (*Recall Bias*). So könnte sich beispielsweise ein Leukämiekranker eher daran erinnern, zeitweise „in der Nähe“ eines Atomkraftwerks gewohnt zu haben als ein Nicht-Erkrankter. Auch ist es vorstellbar, dass Personen, die eine Wirbelsäulenerkrankung haben im Vergleich zu Nicht-Erkrankten berufliche Belastungen ihres Rückens rückblickend als stärker einschätzen.

Als Folge treten Fehlklassifikationen auf, d.h. Personen, welche unter anderen Umständen als Nicht-Exponierte klassifiziert worden wären, werden durch Interviewer oder Recall Bias zu Exponierten. Dies wiederum kann die Ergebnisse verzerren und zu einer Überschätzung des Einflusses des Risikofaktors führen. Während Maßnahmen zur Reduktion des Interviewer Bias bereits erläutert wurden, ist ein Recall Bias nur schwer zu kontrollieren. Er sollte daher als alternative Erklärung der Ergebnisse immer im Hinterkopf behalten werden.

4.3 Störgrößen (Confounding)

Unter Störgrößen versteht man Faktoren, die sowohl mit dem Risikofaktor als auch mit der Erkrankung assoziiert sind. Sie können dadurch einen scheinbaren Zusammenhang zwischen Risikofaktor und Erkrankung erzeugen, während sie selbst Ursache der Erkrankung sind (vgl. Abbildung 5). Ein sehr anschauliches – wenn auch nicht direkt krankheitsbezogenes – Beispiel ist der in verschiedenen Europäischen Ländern beobachtete Zusammenhang zwischen der Storchendichte und der Geburtenrate eines Landes. Trotz der nicht vorhandenen biologischen Plausibilität dieses Zusammenhangs wurde er in mehreren unabhängigen Studien gefunden. Weitete man jedoch seinen Blick und betrachtet Variablen, die sowohl mit der Storchenzahl als auch mit der Geburtenrate in Verbindung stehen, so finden sich durchaus plausible Erklärungen. So kann ein gleichzeitiger Rückgang der Geburtenrate und der Storchendichte mit der parallel verlaufenden Industrialisierung erklärt werden. Vermehrte Trockenlegungen vernichteten die Lebensgrundlage der Störche, während gleichzeitig veränderte Arbeits- und Lebensbedingungen zu einem Rückgang der Geburtenrate führten. Dies macht deutlich, dass parallel verlaufende Entwicklungen noch lange keine Aussagen über kausale *Ursachen* von Ereignissen erlauben.

Im Krankheitskontext könnte Confounding beispielsweise im folgenden Kontext auftreten: In

Untersuchungen erscheint ein bestimmtes Gesundheitsverhalten (z.B. der vermehrte Konsum bestimmter Vitamine) als geeignet für die Vorbeugung einer Krankheit, da es mit einer geringen Erkrankungsrate zusammenhängt. Tatsächlich könnten

jedoch andere Verhaltensweisen, die häufig mit dem vermehrten Konsum von Vitaminen verbunden sind (z. B. das Treiben von Sport) die eigentlichen Faktoren darstellen, die das Auftreten der Erkrankung verhindern.

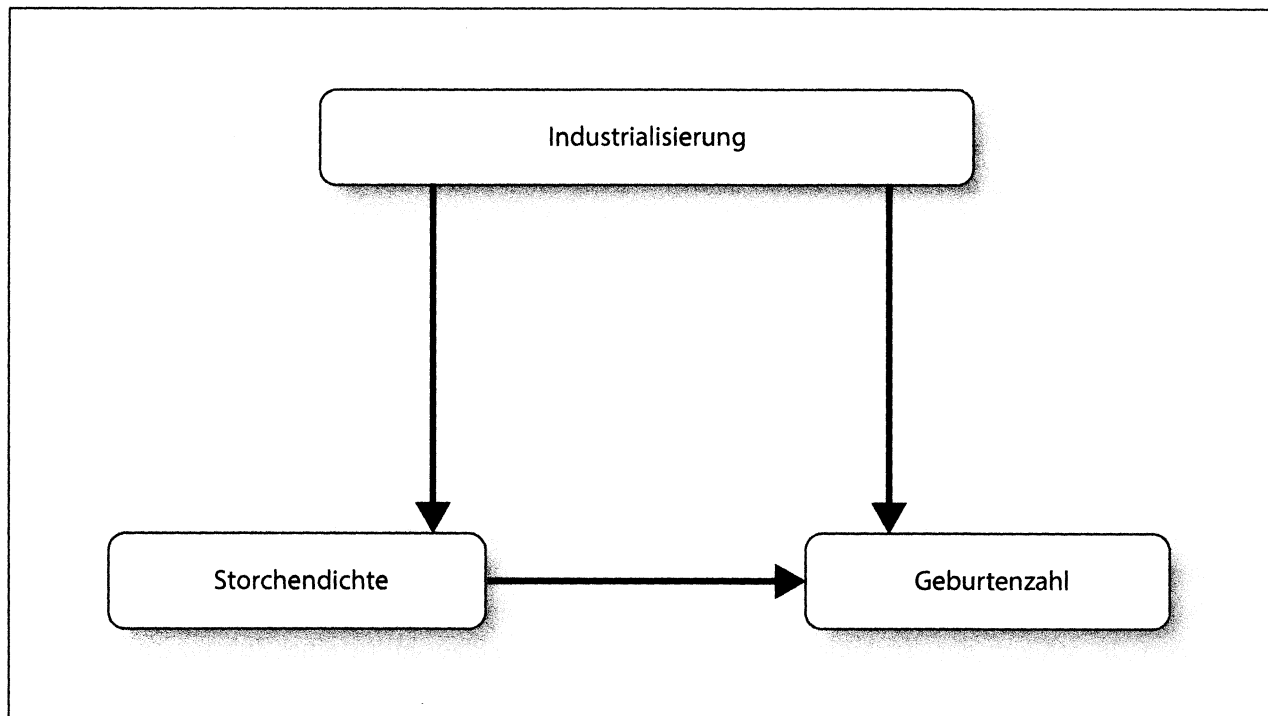


Abbildung 5:
Industrialisierung als Confounder der Beziehung zwischen Storchendichte und Geburtenrate

5 Kriterien zur Beurteilung der Kausalität eines Risikofaktors für eine Krankheit

Wie aus den vorhergehenden Abschnitten deutlich wurde, kann jede Art epidemiologischer Studien verschiedenen Störungen unterliegen. Um Aussagen über die Kausalität von Risikofaktoren für das Auftreten oder den Verlauf einer bestimmten Erkrankung machen zu können, d.h. um die Frage beantworten zu können, ob ein Risikofaktor als Ursache für eine Krankheit angesehen werden kann, sind daher immer mehrere Studien notwendig. Bereits 1965 stellte Bradford Hill mehrere Kriterien vor, die für die Beurteilung von krankheitsbezogenen Risikofaktoren herangezogen werden sollten. Diese können zum Teil auf einzelne, in den meisten Fällen jedoch nur auf die Ergebnisse mehrerer Studien bezogen werden. Sie eignen sich daher vor allem dazu, den Gesamtbeitrag mehrerer Studien zum Zusammenhang eines Risikofaktors mit einer Krankheit zu beurteilen.

Stärke: Je stärker der Zusammenhang von Risikofaktor und Krankheit, desto wahrscheinlicher ist es, dass beide kausal verbunden sind. Dieses Kriterium lässt sich sowohl auf eine einzelne Studie beziehen als auch auf den mittleren Zusammenhang zwischen Risikofaktor und Krankheit über mehrere Studien hinweg.

Konsistenz: Je übereinstimmender die Ergebnisse mehrerer Studien, umso wahrscheinlicher ist eine kausale Beziehung. Dies gilt umso mehr, je unterschiedlicher die Studien sind, da sich dadurch die Wahrscheinlichkeit verringert, dass ein Effekt durch dieselben Störfaktoren oder methodischen Einschränkungen verursacht ist, ohne dass eine Kausalbeziehung zwischen Risikofaktor und Krankheit vorliegt.

Spezifität: Ein Risikofaktor, der nur mit einer bestimmten Erkrankung zusammenhängt, ist mit einer größeren Wahrscheinlichkeit auch eine Ursache dieser Erkrankung als ein Faktor, der mit vielen verschiedenen Krankheiten assoziiert ist.

Zeitlichkeit: Stärkere Belege für den Kausalzusammenhang zwischen bestimmten Faktoren und einer Erkrankung liefern solche Studien, die eindeutig zeigen können, dass der Risikofaktor *vor* dem Auftreten der Krankheit bestand. Dies ist bei ökologischen Prävalenzstudien nicht, bei Fall-Kontroll-Studien nur bedingt gegeben. Eindeutige Hinweise auf die Zeitlichkeit von Risikofaktor und Erkrankung liefern vor allem Kohorten- und Interventionsstudien.

Biologischer Gradient: Häufig kann davon ausgegangen werden, dass kausale Faktoren einem gewissen biologischen Gradienten unterliegen, d.h. je stärker der Risikofaktor ausgeprägt ist (z. B. Anzahl von Zigaretten pro Tag), umso häufiger sollte eine bestimmte Erkrankung (z. B. Lungenkrebs) auftreten.

Biologische Plausibilität: Selbst wenn die meisten anderen Kriterien erfüllt sind, sollte immer die biologische Plausibilität eines Zusammenhangs hinterfragt werden. So kann man für die weiter oben beschriebene Beziehung zwischen Storchenzahl und Geburtenrate nur schwer eine sinnvolle biologische Erklärung finden. In einem solchen Fall ist dies als ein starker Hinweis darauf zu werten, dass der beobachtete Zusammenhang eher nicht kausaler Natur ist.

Analogie: Zusätzliche Hinweise auf die Kausalität einer Beziehung zwischen Risikofaktor und Krankheit können auch durch eine Ausweitung des Blicks auf Studien mit ähnlichen Erkrankungen gefunden werden. Finden sich dort ähnliche Risikofaktoren (idealerweise jedoch nicht die gleichen, vgl. Spezifität), so ergeben sich daraus weitere Argumente dafür, dass der vermutete Risikofaktor tatsächlich die bzw. eine Ursache für die untersuchte Erkrankung darstellt.

Experimentelle Bestätigung: Die stärksten Kausalitätshinweise liefern Ergebnisse aus Studien, in denen der Risikofaktor experimentell beeinflusst

wurde. Dabei sollten Interventionen durchgeführt werden, die dem vermuteten Wirkmechanismus des Risikofaktors entsprechen. Experimentell bedeutet weiterhin, dass es nicht nur eine Interventionsgruppe, sondern auch eine Kontrollgruppe ohne Intervention gibt, wobei Patienten möglichst unsystematisch (zufällig) einer von beiden Gruppen zugewiesen werden sollten. Ändert sich daraufhin die Erkrankungsrate bzw. der Krankheitsverlauf der Interventionsgruppe im Vergleich zur Kontrollgruppe, so kann der kausale Zusammenhang zwischen Risikofaktor und Erkrankung als experimentell bestätigt gelten.

Häufig ist es nicht möglich, v. a. während eines Forschungsprozesses, alle genannten Kriterien zu

überprüfen oder zu bestätigen. Je nach Forschungsstand gibt es möglicherweise noch keine ausreichende Anzahl von Studien, um Konsistenz, Spezifität oder Analogien zu beurteilen. Gleichzeitig müssen Modelle zu biologischen Wirkmechanismen möglicherweise erst noch entwickelt werden. Auch die experimentelle Bestätigung ist teilweise (z. B. aus ethischen Gründen) nicht möglich oder kann erst relativ spät am Ende eines Forschungsprozesses durchgeführt werden. Daher kann als Leitlinie für die Beurteilung des Forschungsstands angesehen werden, dass die Wahrscheinlichkeit einer kausalen Beziehung zwischen Risikofaktor und Krankheit umso höher ist, je mehr der genannten Kriterien erfüllt sind.

6 Die Deutsche Wirbelsäulenstudie

Abschließend sollen auf dem Hintergrund der ausgeführten Informationen Anregungen für die Einordnung der Ergebnisse der Deutschen Wirbelsäulenstudie gegeben werden. Zu diesem Zweck erfolgt zunächst ein Bezug des Studiendesigns und der Stichprobe auf die dargestellten Informationen. Im Anschluss werden Hinweise zur Beurteilung des Forschungsstands anhand der Bradford-Hill-Kriterien, insbesondere (aber nicht ausschließlich) aufgrund der Ergebnisse der Deutschen Wirbelsäulenstudie gegeben.

6.1 Studiendesign

Bei der Deutschen Wirbelsäulenstudie handelt es sich um eine Fall-Kontroll-Studie, bei der eine Gruppe von Erkrankten mit einer etwa gleich großen Gruppe Nicht-Erkrankter hinsichtlich ihrer (retrospektiv erhobenen) Risikofaktoren verglichen wird. Im Fokus stand hierbei nicht ein einzelner Risikofaktor, sondern das Zusammenwirken verschiedener Belastungen auf die Entwicklung einer Erkrankung der LWS. Nach Röhrig et al. (2009: 265) sind Fall-Kontroll-Studien besonders dazu geeignet, Auswirkungen multipler Expositionen zu untersuchen, was für die Angemessenheit der Auswahl des Untersuchungsdesigns, für die Fragestellung spricht. Dennoch sind auch mit diesem Studiendesign gewisse Nachteile verbunden (vgl. Abschnitt 3.3). Maßnahmen zur Reduzierung dieser Nachteile werden in den folgenden beiden Abschnitten diskutiert.

6.2 Stichprobe

Es erfolgte eine multizentrische Rekrutierung der Teilnehmer, d.h. die Teilnehmer kamen aus verschiedenen Kliniken und Einzugsgebieten (Frankfurt am Main, Freiburg, Halle, Regensburg). Damit

wurde die Wahrscheinlichkeit reduziert, dass die gefundenen Zusammenhänge nur auf eine spezifische Patientengruppe zutreffen, die in einer ganz bestimmten Klinik eines einzelnen Einzugsgebiets anzutreffen sind. Durch die Auswahl etwa gleich vieler männlicher und weiblicher Teilnehmer und das Einbeziehen zweier verschiedener Erkrankungen (Prolaps, Chondrose) wurde einerseits die Untersuchung von Grenzwerten sowohl für männliche als auch für weibliche Patienten und andererseits der Vergleich der differenziellen Auswirkungen von Belastungen auf zwei verschiedene Erkrankungsarten ermöglicht. Mit einer Fallzahl von jeweils über 900 Personen in Patienten- und Kontrollgruppe sowie von immer noch 150 bis 280 Patienten pro Fallgruppe kann von einer ausreichend großen Stichprobengröße für die berechneten statistischen Analysen ausgegangen werden (vgl. u. a. Peng et al. 2002: 10). Entgegen den Empfehlungen für Fall-Kontroll-Studien wurde von einem direkten Matching der in die Patienten- und in die Kontrollgruppe aufgenommenen Teilnehmer abgesehen, stattdessen wurde als Kontrollstichprobe eine Zufallsauswahl der Bevölkerung der entsprechenden Einzugsgebiete herangezogen. Dies hatte jedoch den Vorteil, dass dadurch eine Schätzung der Inzidenzrate der beiden Erkrankungen möglich wurde. Entsprechende Zahlen lagen bisher noch nicht vor, waren jedoch als Information für die Interpretation der Ergebnisse von Bedeutung.

Wie bereits in Abschnitt 2 angesprochen, kann bei Fall-Kontroll-Studien kein Relatives Risiko berechnet werden, da die Randverteilungen nicht zufällig sind (die Anzahl an Teilnehmern mit und ohne Erkrankung ist bei diesem Design etwa gleich groß und entspricht damit nicht der tatsächlichen Verteilung der Bevölkerung). Stattdessen werden Odds Ratios berechnet und als Schätzungen für das Erkrankungsrisiko herangezogen. Diese entsprechen jedoch nur bei einer geringen Inzidenzrate in etwa dem Relativen Risiko. Die Deutsche Wirbelsäulen-

lenstudie ermittelte anhand der Kontrollstichprobe Prävalenzraten von 0,2 bis 0,4% für Prolapse und 0,1 bis 0,3% für Chondrosen (Seidler et al. 2012: 32), womit vermutlich auch die Inzidenzraten Nicht-Exponierter die geforderten 1% (vgl. Abschnitt 2) deutlich unterschreiten. Somit kann davon ausgegangen werden, dass die für die weiteren Berechnungen zugrunde gelegten Odds Ratios von 2 als Maß für die Verdoppelungsrate brauchbare Schätzungen für das Relative Risiko darstellen.

Obwohl aus den beschriebenen Gründen von einem direkten Matching abgesehen wurde, wurde auf eine möglichst hohe Vergleichbarkeit von Kontroll- und Patientenstichprobe auf verschiedenen relevanten Variablen geachtet. So lag die Altersspanne bei beiden zwischen 25 und 70 Jahren, mit vergleichbaren Altersmedianen² zwischen 46 und 48 Jahren. Zwar war die Teilnahmerate bei den Erkrankten mit 66,4% etwas höher als in der Bevölkerungsstichprobe (53, 4%) und in beiden Gruppen nahmen mehr Personen aus Dienstleistungs- als aus Arbeiter- und Handwerksberufen teil, systematische Unterschiede der Teilnahmerate abhängig vom Sozialstatus konnten jedoch ausgeschlossen werden. Damit wurde zumindest im Hinblick auf den sozialen Status ein verfälschender Response Bias überprüft und ausgeschlossen. Weitere relevante Variablen wie Größe, Gewicht, Belastungen durch Freizeitaktivitäten oder Vorerkrankungen (vollständige Liste vgl. Seidler et al. 2007: 294) wurden als Confounder (Kontrollvariablen) in die Regressionsgleichung einbezogen, um einen möglichen Einfluss anderer Variablen als der im Interesse stehenden beruflichen Wirbelsäulenbelastungen zu kontrollieren.

6.3 Datenerhebung

Risikofaktoren: Die Erhebung der Risikofaktoren erfolgte mittels eines persönlichen 1,5-stündigen

Interviews zu Wirbelsäulenbelastungen durch Lastenhandhabung, Rumpfbeugung und Ganzkörperschwingungen (Erstinterview) sowie – bei Personen, die in diesem Interview einen bestimmten Grenzwert überschritten – mittels eines Experteninterviews zur detaillierteren Beschreibung der Belastungen (Zweitinterview) durch Mitarbeiter der Technischen Aufsichtsdienste (TAD) der gesetzlichen Unfallversicherungen. Bei den Erstinterviewern handelte es sich um geschulte Laien-Interviewer, die auf einen standardisierten Interviewleitfaden zurückgriffen. In diesem waren sowohl der Fragentext als auch die Antwortvorgabe festgelegt. Bei der Schulung eigneten sich die Interviewer wichtige Regeln an, die sicherstellen sollten, dass alle Befragten möglichst auf dieselbe Art und Weise befragt wurden. So sollte eine „professionelle Neutralität gegenüber dem Befragten“ unter anderem dadurch hergestellt werden, dass das Interview nicht in Form eines Gesprächs durchgeführt, sondern immer eine „Frage-Antwort-Situation“ aufrechterhalten wurde (Linhardt et al. 2007: 246). Der Interviewer durfte lediglich bei der Klärung von Fragen unterstützen. Auch die Eingangsphase des Interviews erfolgte anhand einer standardisierten Einführung, die das Ziel der Untersuchung darstellte, ohne auf die Hauptfragestellung einzugehen („Wir führen eine Studie zum Thema Arbeit und Gesundheit durch und wollen mit der Befragung die beruflichen Belastungen abklären“). Darüber hinaus wurde Verzerrungen durch die Interviewer vorgebeugt, indem ihnen nicht mitgeteilt wurde, welcher der beiden Gruppen ihr Gegenüber angehörte.

Die zweite Phase der Erhebung wurde anhand von arbeitstechnischen Fragebögen durchgeführt. Abgesehen von dem Hinweis auf die hohe fachliche (und in den meisten Fällen auch berufsspezifische) Kompetenz der Interviewer enthalten die Berichte über diesen Teil der Studie keine konkreteren Informationen über Schulungen und Instruktionen dieser Mitarbeiter; Interviewer-Einflüsse wurden aber auch hier weitgehend reduziert, indem den Mitarbeitern bei der Befragung nicht bekannt war, ob sie einen Patienten oder eine Kontrollperson interviewten.

2 Alter, das die Probanden in zwei Hälften teilt, d.h. 50% der Probanden waren jünger oder gleich alt, 50% älter oder gleich alt.

Die Komplexität der Interviews zur Abschätzung der Risikovariablen zeigt sich jedoch unter anderem darin, dass die Übereinstimmungen der Klassifikation von Probanden, die für zwei der betrachteten Dosismodelle überprüft wurden, in einigen Fällen nicht als ausreichend angesehen werden konnten. So ergaben sich für die Übereinstimmung innerhalb der Interviewer (Interview eines Probanden wurde von demselben Interviewer zweimal durchgeführt, wobei das Zeitintervall zwischen beiden Interviews nicht angegeben wurde) noch akzeptable Kappas³ um .65. Bei der Übereinstimmung zwischen Interviewern (ein Proband wurde von zwei verschiedenen Interviewern befragt, Zeitintervall ca. ein Jahr) variierten die Werte mit Kappas zwischen .12 und .68 jedoch deutlich stärker und waren damit in einigen Fällen nicht als ausreichend zu beurteilen (Ellegast et al. 2007: 262).

Abgesehen von den Vorkehrungen zur Reduzierung des Interviewer Bias bleibt das in der Natur retrospektiver Studien liegende Problem des Erinnerungsbias, indem Belastungen der Wirbelsäule auf vergangenen Arbeitsstellen von Wirbelsäulenerkrankten möglicherweise anders (d. h. wahrscheinlich intensiver) erinnert werden als von Nicht-Erkrankten. Dies könnte nur durch eine prospektive Studie ausgeschlossen bzw. überprüft werden.

Diagnose: Die Diagnosestellung als Grundlage für die Rekrutierung der Teilnehmer der Fallgruppe erfolgte aufgrund der klinischen Untersuchung sowie auf der Basis radiologischer Befunde durch den niedergelassenen oder den Klinikradiologen. Die infrage kommenden Patienten wurden nur dann in die Fallgruppe aufgenommen, wenn ihre

radiologischen Befunde durch einen zweiten, von der Studienleitung beauftragten Radiologen bestätigt wurden. Dabei zog dieser die entsprechenden krankheitsspezifischen standardisierten Kriterien der gewerblichen Berufsgenossenschaften zur Begutachtung der jeweiligen Berufskrankheit (Bolm-Audorff et al. 2005) heran. Zusätzlich erfolgte eine stichprobenhafte Überprüfung der Zweitgutachten durch einen Drittgutachter (Übereinstimmung: Kappa-Werte von .80 für Prolaps und von .86 für Chondrose, was für eine sehr gute, wenn auch nicht perfekte Übereinstimmung spricht).

Als Kriterium für die Fallrekrutierung wurde festgelegt, dass die Erstmanifestation der Krankheit nicht mehr als zehn Jahre zurückliegen durfte (vgl. Linhardt et al. 2007: 246). Darüber hinaus wurden nur solche Belastungen im Interview erfasst, die vor der Erstmanifestation auftraten (vgl. Seidler et al. 2007: 293). Dadurch sollte verhindert werden, dass sich das Auftreten der betrachteten Risikofaktoren und der Erkrankung zeitlich überschneiden. Zwar betrug bei 50% der Fälle die Differenz zwischen Erstmanifestation und Interview maximal vier Monate, bei Fällen, deren Erstmanifestation (und damit auch die Erinnerung an zuvor aufgetretene Belastungen) mehrere Jahre zurücklag, könnte jedoch eine stärkere Verzerrung durch Erinnerungseffekte aufgetreten sein. Das Vorhandensein systematischer Verzerrungen durch Erinnerungsfehler zwischen Patienten- und Kontrollgruppe kann somit nicht vollständig ausgeschlossen werden, da bei den Kontrollpersonen alle Belastungen bis zum Zeitpunkt der Befragung erfasst wurden und somit solche Erinnerungsverzerrungen in dieser Gruppe weniger ins Gewicht fielen.

3 Maß für die Übereinstimmung zwischen Beurteilern; Wertebereich zw. -1 und 1, wobei Werte von .75 bis 1 eine sehr gute, von .40 bis .75 eine akzeptable und Werte unter .40 eine unzureichende Übereinstimmung der Beurteiler kennzeichnen.

7 Einordnung des mit der Deutschen Wirbelsäulenstudie erreichten Forschungsstands anhand der Bradford-Hill-Kriterien

Wie bereits bei der Beschreibung der Bradford-Hill-Kriterien im Abschnitt 5 angesprochen, können diese nur zum Teil für die Beurteilung einzelner Studien herangezogen werden. Bezieht man jedoch auch die Ergebnisse vorheriger Studien mit ein, so kann zumindest der Beitrag der einzelnen Studien zum gesamten Forschungsstand im Hinblick auf die einzelnen Kriterien abgeschätzt werden. Im Folgenden werden einige Hinweise zur Beurteilung des Forschungsstands zum Zusammenhang zwischen beruflichen Wirbelsäulenbelastungen und Erkrankungen der LWS in Bezug auf die Bradford-Hill-Kriterien unter der besonderen Berücksichtigung der Ergebnisse der Deutschen Wirbelsäulenstudie gegeben.

Stärke: Eine ausreichende Stärke des Zusammenhangs der ermittelten Grenzwerte mit den untersuchten Erkrankungen wurde mit Bezug auf die Verdopplungsdosis mittels der Forderung eines Odds Ratios von mindestens 2 sichergestellt.

Konsistenz: Die Deutsche Wirbelsäulenstudie setzt vorhergehende Studien zum Zusammenhang zwischen Wirbelsäulenbelastungen und Erkrankungen der Lendenwirbelsäule fort und bestätigt diese. Wie die meisten wissenschaftlichen Studien strebt jedoch auch sie neue Erkenntnisse mit differenzierteren Erhebungsmethoden und exakteren statistischen Modellen an und liefert diese. Die Konsistenz dieser neuen, über die Erkenntnisse vorheriger Studien hinausgehenden Ergebnisse muss naturgemäß erst in weiteren Studien überprüft werden.

Spezifität: Die Modelle wurden für die beiden Erkrankungen Prolaps und Chondrose getrennt überprüft. Dies sollte spezifische Vorhersagen für beide Erkrankungen ermöglichen. Obwohl es einzelne Unterschiede in der Anpassung verschiedener Modelle für die beiden Erkrankungsarten gab, konnte diese Spezifität statistisch jedoch nicht bestätigt werden (Seidler et al. 2007:301).

Zeitlichkeit: Der zeitlichen Abfolge von Risikofaktoren (Wirbelsäulenbelastungen) und Erkrankung wurde in der DWS insofern Rechnung getragen, als nur solche Belastungen erhoben wurden, die vor der Erkrankung auftraten. Da bei der Mehrzahl der Patienten Erstdiagnose und Erhebungszeitpunkt relativ nah beieinander lagen, sollte dies in den überwiegenden Fällen gelungen sein. Bei Patienten, deren Erstdiagnose dagegen bereits deutlich länger (mehrere Jahre) zurücklag, kann nicht mit vollständiger Sicherheit davon ausgegangen werden, dass die genannten Belastungen auch tatsächlich vor der Erkrankung aufgetreten waren. Zudem könnten Verzerrungen der Erinnerung hinsichtlich der Stärke der Belastung aufgetreten sein. Diese Erinnerungsfehler sollten jedoch bei der Größe der befragten Stichprobe die Ergebnisse nur bedingt beeinflusst haben.

Biologischer Gradient: Ein Ziel der Deutschen Wirbelsäulenstudie war die Entwicklung alternativer (besserer) Modelle zur Abschätzung des Einflusses kumulativer Wirbelsäulenbelastungen auf spätere Erkrankungen der LWS. Bei der Beurteilung der Güte der verschiedenen getesteten Modelle wurde dabei solchen der Vorzug gegeben, die einen monotonen Anstieg der Dosis-Wirkungsbeziehung zeigten. Damit wurde der Forderung eines biologischen Gradienten Rechnung getragen. Dennoch zeigten einige der ansonsten gut angepassten Modelle einen Abfall der Vorhersagegüte im Bereich sehr hoher Belastung. Die Autoren vermuten, dass dieser Abfall durch einen Healthy-Worker-Effekt erklärt werden könnte. Personen mit starken Wirbelsäulenbeschwerden könnten in der Gruppe der Hochbelasteten unterrepräsentiert sein, indem sie bereits vor der Diagnose ihrer Erkrankung nicht mehr in der Lage waren, die entsprechenden Tätigkeiten auszuführen, ihren Arbeitsplatz wechselten und somit ihre als Risikofaktor in die Studie eingehenden Belastungen bereits vor

Beginn der Studie reduziert hatten (vgl. Seidler et al. 2007: 300). Ob diese Hypothese geeignet ist, den beobachteten Effekt plausibel und v.a. vollständig zu erklären, kann jedoch diskutiert werden.

Biologische Plausibilität: Auf das Kriterium biologischer Plausibilität wird explizit erst bei dem Vergleich statistisch gleich guter Modelle eingegangen. Hierbei werden zunächst solche Modelle bevorzugt, die für Männer und Frauen gleichermaßen zutreffen. Erst in einem zweiten Schritt werden die biologisch plausibleren Modelle präferiert.

Dennoch sollte davon ausgegangen werden, dass bei der Auswahl der getesteten Modelle prinzipiell nur solche herangezogen wurden, denen eine gewisse biologische Plausibilität zugeschrieben werden kann. Um dies im Einzelnen zu beantworten, sind jedoch spezifische medizinische, biomechanische und statistische Kenntnisse notwendig.

Experimentelle Bestätigung: Eine experimentelle Bestätigung der Ergebnisse ist aufgrund der Komplexität der Modelle sowie aufgrund ethischer Probleme nur bedingt möglich.

8 Fazit

Ein wesentliches Anliegen der Deutschen Wirbelsäulenstudie war eine detaillierte(re) Erkrankungs- und Belastungsbeschreibung mit dem Ziel der Verbesserung von Ansätzen zur Prüfung von Dosis-Wirkungsbeziehungen zwischen Lumbalbelastungen und Erkrankungen der Lendenwirbelsäule. Dazu wurden an einer Stichprobe von über 1.000 Personen umfassende Interviews über berufliche Belastungen durchgeführt, diese biomechanisch analysiert und mit Hilfe verschiedener Modelle kumulative Lebenszeitdosen modelliert. Der detaillierten Beschreibung von Belastungen sowie deren differenzierten Berücksichtigung bei der Berechnung der Lebenszeitdosis mit Hilfe verschiedener Modelle wurde damit in der Studie umfassend Rechnung getragen.

Diese tiefer gehende Betrachtung ist jedoch

naturgemäß mit einer hohen Komplexität der Studie in Bezug auf die Datenerhebung, -analyse und -interpretation verbunden. Entsprechend wird von den Autoren angemerkt, dass die „Genauigkeit der Belastungsprognosen durch die biomechanischen Modellrechnungen ... wesentlich von den Modellannahmen sowie insbesondere auch von der Genauigkeit der erhobenen Eingangsdaten zur Beschreibung der externen Belastungsfaktoren [abhängt]“ (Jäger et al. 2007, 273). Für die Beurteilung der Sorgfalt bei der Erhebung der Eingangsdaten werden ausführliche Informationen geliefert, so dass diese mit entsprechendem methodischen Wissen in ausreichendem Maße möglich ist. Hierzu sollte der vorliegende Beitrag einige, hoffentlich auch für den (methodischen) Laien verständliche Hintergrundinformationen liefern.

Literatur

- BOLM-AUDORFF U et al. (2005) Medizinische Beurteilungskriterien zu bandscheibenbedingten Berufskrankheiten der Lendenwirbelsäule (I). In: Trauma Berufskrankheit 7: 211 – 252.
- ELLEGASTR et al. (2007) Erhebungen zur beruflichen Wirbelsäulenexposition durch die Technischen Aufsichtsdienste der Unfallversicherungsträger im Rahmen der Deutschen Wirbelsäulenstudie. In: Zentralblatt für Arbeitsmedizin, Arbeitsschutz und Ergonomie: 251 – 263.
- FLETCHER RH, FLETCHER SW (2007) Klinische Epidemiologie, Bern.
- HILL AB (1965) The environment and disease: Association or causation? In: Proceedings of the Royal Society of Medicine 58: 295 – 300.
- JÄGER M et al. (2007) Biomechanische Analysen zur Belastung der Lendenwirbelsäule innerhalb der Deutschen Wirbelsäulenstudie. In: Zentralblatt für Arbeitsmedizin, Arbeitsschutz und Ergonomie 57: 264 – 276.
- LINHARDT O et al. (2007) Studiendesign der Deutschen Wirbelsäulenstudie. In: Zentralblatt für Arbeitsmedizin, Arbeitsschutz und Ergonomie: 243 – 250.
- PENG CJ et al. (2002) An introduction to logistic regression analysis and reporting. In: The Journal of Educational Research 96: 3 – 14.
- RÖHRIG B et al. (2009) Studientypen in der medizinischen Forschung. In: Deutsches Ärzteblatt 106: 262 – 268.
- SEIDLER A et al. (2007) Zusammenhang zwischen der kumulativen Wirbelsäulenbelastung durch Lastenhandhabungen und lumbalen Prolapserkrankungen – Ergebnisse der Deutschen Wirbelsäulenstudie. In: Zentralblatt für Arbeitsmedizin, Arbeitsschutz und Ergonomie 57: 290 – 303.
- SEIDLER A et al. (2012) Abschlussbericht zum Forschungsvorhaben „Erweiterte Auswertung der Deutschen Wirbelsäulenstudie mit dem Ziel der Ableitung geeigneter Richtwerte“, Dresden.
- WOLF-MAIER K et al. (2003) Hypertension prevalence and blood pressure levels in 6 European countries, Canada, and the United States. In: The Journal of the American Medical Association 289: 2363 – 2369.